ISSN:2828-6235
Impact factor:13.77

International Journal of Literature and Language Studies(IJLLS)

LINGUISTIC AND TECHNOLOGICAL FOUNDATIONS FOR BUILDING A LARGE-SCALE UZBEK TEXT CORPUS FOR NATURAL LANGUAGE PROCESSING

O‘ralova Feruza O‘ktam qizi
University of Exact and Social Sciences
uralovaferuzauktamqizi@gmail.com

Abstract. This article investigates the creation of a large-scale Uzbek text corpus for Natural Language Processing (NLP) systems within the implementation framework of Presidential Decree No. UP-6084 (October 20, 2020) and Presidential Resolutions No. PP-358 (October 14, 2024) and No. PP-308 (October 21, 2025). The study analyzes the impact of the agglutinative nature of the Uzbek language, complex graphemes of the current Latin alphabet (digraphs *sh, ch, ng* and apostrophe-modified letters *o‘, g‘*), and the parallel use of Cyrillic and Latin scripts on computational linguistics processes. The author proposes a corpus architecture encompassing text collection, orthographic normalization in Unicode (NFC), contextual transliteration, lemmatization, and morphological tagging. Based on the analysis of domestic and international computational linguistics research, actionable recommendations are formulated for developing a high-precision Uzbek text corpus.

Key words: natural language processing (NLP), text corpus, computational linguistics, agglutinative language, Unicode normalization, Latin alphabet, lemmatization, morphological tagging.

TABIIY TILNI QAYTA ISHLASH UCHUN KENG KO‘LAMLI O‘ZBEK TILIDAGI MATNLAR KORPUSINI YARATISHNING LINGVISTIK VA TEXNOLOGIK ASOSLARI

Annotatsiya. Ushbu maqolada O‘zbekiston Respublikasi Prezidentining 2020-yil 20-oktabrdagi PF–6084-son Farmoni, 2024-yil 14-oktabrdagi PQ–358-son hamda 2025-yil 21-oktabrdagi PQ–308-son qarorlari ijrosi doirasida tabiiy tilni qayta ishlash (NLP) tizimlari uchun keng ko‘lamli o‘zbek tilidagi matnlar korpusini yaratish masalalari tadqiq etilgan. Tadqiqotda o‘zbek tilining agglyutinativ tabiati, amaldagi lotin alifbosidagi murakkab grafemalar (digraflar: *sh, ch, ng* va apostrofli *o‘, g‘* harflari) hamda kirill-lotin grafikalari parallel qo‘llanilishining kompyuter lingvistikasi jarayonlariga ta’siri tahlil qilingan. Muallif matnlarni yig‘ish, Unicode (NFC) standartida orfografik normalizatsiya qilish, kontekstual transliteratsiya, lemmatizatsiya va morfologik teglash bosqichlarini o‘z ichiga olgan korpus arxitekturasini taklif etadi. Mahalliy va xorijiy kompyuter lingvistikasi olimlarining ilmiy ishlarini tahlil qilish asosida yuqori aniqlikdagi o‘zbek tili matnlar korpusini shakllantirish bo‘yicha amaliy tavsiyalar ishlab chiqilgan.

Kalit so‘zlar: tabiiy tilni qayta ishlash (NLP), matnlar korpusi, kompyuter lingvistikasi, agglyutinativ til, Unicode normalizatsiya, lotin alifbosi, lemmatizatsiya, morfologik teglash.

ЛИНГВИСТИЧЕСКИЕ И ТЕХНОЛОГИЧЕСКИЕ ОСНОВЫ СОЗДАНИЯ КРУПНОМАСШТАБНОГО КОРПУСА ТЕКСТОВ НА УЗБЕКСКОМ ЯЗЫКЕ ДЛЯ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Аннотация. В данной статье исследуются вопросы создания крупномасштабного корпуса текстов на узбекском языке для систем обработки естественного языка (NLP) в рамках реализации Указа Президента Республики Узбекистан № УП-6084 от 20 октября 2020 года, а также Постановлений № ПП-358 от 14 октября 2024 года и № ПП-308 от 21 октября 2025 года. Проанализировано влияние агглютинативной природы узбекского языка, сложных графем действующего латинского алфавита (диграфов *sh, ch, ng* и букв с апострофом *o‘, g‘*)

и параллельного использования кириллицы и латиницы на процессы компьютерной лингвистики. Автор предлагает архитектуру корпуса, включающую сбор текстов, орфографическую нормализацию в стандарте Unicode (NFC), контекстную транслитерацию, лемматизацию и морфологическую разметку. На основе анализа трудов отечественных и зарубежных ученых разработаны практические рекомендации по формированию высокоточного корпуса текстов узбекского языка.

Ключевые слова: обработка естественного языка (NLP), корпус текстов, компьютерная лингвистика, агглютинативный язык, нормализация Unicode, латинский алфавит, лемматизация, морфологическая разметка.

Kirish. Zamonaviy sun’iy intellekt texnologiyalari va tabiiy tilni qayta ishlash (Natural Language Processing — NLP) tizimlarining jadal taraqqiyoti milliy tillarni raqamlashtirish va ularning elektron ma’lumotlar bazasini shakllantirishni strategik vazifaga aylantirdi. O‘zbekiston Respublikasi Prezidentining 2020-yil 20-oktabrdagi "Mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to‘g‘risida"gi PF–6084-son Farmoni, 2024-yil 14-oktabrdagi "Sun’iy intellekt texnologiyalarini 2030-yilgacha rivojlantirish strategiyasini tasdiqlash to‘g‘risida"gi PQ–358-son hamda 2025-yil 21-oktabrdagi "Davlat tilini yanada rivojlantirishga oid qo‘shimcha chora-tadbirlar to‘g‘risida"gi PQ–308-son qarorlari o‘zbek tilini zamonaviy axborot texnologiyalari va sun’iy intellekt modellariga keng integratsiya qilish, shuningdek, O‘zbek tilining milliy korpusini yaratish uchun mustahkam huquqiy-tashkiliy asos yaratdi. Katta til modellari (LLM), mashina tarjimasi va nutqni qayta ishlash tizimlarining samaradorligi bevosita keng ko‘lamli, janr jihatidan muvozanatli va orfografik toza matnlar korpusiga tayanadi. Biroq amaliyotda o‘zbek tilidagi matnlar korpusini yaratishda bir qator lingvistik va texnologik murakkabliklar mavjud. Xususan, o‘zbek tilining agglyutinativ grammatik tuzilishi, bitta o‘zakka ko‘plab affikslarning ketma-ket qo‘shilishi, amaldagi 1995-yilgi lotin alifbosidagi digraflar (*sh, ch, ng*) va apostrofli (*o‘, g‘*) harflarning turlicha kodlanishi, shuningdek, raqamli makonda ham kirill, ham lotin grafikalaridagi matnlarning parallel mavjudligi avtomatik qayta ishlash algoritmlariga jiddiy ta’sir ko‘rsatadi. Ushbu maqolaning maqsadi tabiiy tilni qayta ishlash vazifalari uchun keng ko‘lamli o‘zbek tili matnlar korpusini yaratish bosqichlarini tizimlashtirish, amaldagi alifbo grafikasidagi texnik muammolarni tahlil qilish va korpus arxitekturasini takomillashtirish bo‘yicha ilmiy asoslangan yechimlarni taklif etishdan iborat.

Mavzuga doir adabiyotlar tahlili. Korpus lingvistikasi va tabiiy tilni avtomatik qayta ishlashning nazariy hamda amaliy masalalari jahon va o‘zbek olimlari tomonidan izchil o‘rganilmoqda. Xorijiy kompyuter lingvistikasida J. Sinkler korpusni muayyan lingvistik mezonlar asosida tizimli ravishda yig‘ilgan elektron matnlar majmuasi sifatida ta’riflab, uning reprezentativlik (tilning barcha uslubiy qatlamlarini mutanosib qamrab olish) tamoyilini asoslab bergan [1, 18-b.]. Tabiiy tilni qayta ishlash algoritmlari, sub-word tokenizatsiya (BPE, WordPiece) va morfologik tahlil modellarining dasturiy asoslari D. Jurafskiy va J. Martinlarning fundamental tadqiqotlarida mukammal yoritilgan [2, 46-b.]. T. Makeneri va E. Uilsonlar esa korpuslarni lingvistik teglash (annotatsiya qilish) standartlari va orfografik birlikning ma’lumotlar bazasi aniqligiga ta’sirini tizimli tahlil qilganlar [3, 68-b.]. O‘zbek tilshunosligida lotin yozuviga asoslangan o‘zbek alifbosi va imlo qoidalarini takomillashtirish, shuningdek, raqamli muhitda yozuv birligini ta’minlash masalalari N. Mahmudov, A. Madvaliyev va N. Mahkamovlarning tadqiqotlarida ilmiy jihatdan asoslangan [4, 14-b.]. O‘zbek kompyuter lingvistikasi maktabi doirasida o‘zbek tilining matematik modellarini yaratish, matnlar bazasini shakllantirish va qidiruv tizimlarini ishlab chiqish tamoyillari A. Po‘latov asarlarida chuqur yoritilgan [5, 118-b.]. Zamonaviy o‘zbek korpus lingvistikasida agglyutinativ til xususiyatlarini hisobga olgan holda morfologik analizatorlar, lemmatizatorlar va elektron lug‘atlar yaratish texnologiyalari N. Abdurahmonova tadqiqotlarida o‘z aksini topgan [6,

84-b.]. O‘zbek tili korpuslarida grafik variantdorlik, leksik birliklarni standartlashtirish va semantik teglash muammolari Sh. Hamroyeva ishlarida o‘rganilgan [7, 59-b.]. Milliy korpusni shakllantirishning nazariy-metodologik asoslari va janrlar muvozanatini ta’minlash masalalari B. Mengliyev hamda R. Karimovlar tomonidan tadqiq etilgan [8, 88-b.]. Shuningdek, o‘zbek tilida keng ko‘lamli NLP ma’lumotlar to‘plamlarini (dataset) yaratish va neyron tarmoq modellarini o‘qitish tajribalari M. Sharipov va E. Kuriyozovlarning xalqaro ilmiy maqolalarida bayon etilgan [9, 136-b.]. Kirill va lotin yozuvlari o‘rtasidagi kontekstual transliteratsiya hamda Unicode normalizatsiya masalalari B. Elov va Sh. Hamroyevalar tadqiqotlarida tahlil qilingan [10, 42-b.]. Shunga qaramasdan, PQ–308-son va PQ–358-son qarorlar talablaridan kelib chiqib, kirill va amaldagi lotin grafikalari asosidagi matnlarni yagona NLP korpusiga birlashtiruvchi yaxlit normalizatsiya va teglash arxitekturasini takomillashtirish masalasi tizimli tadqiqotni talab etadi.

Tadqiqot obyekti va qo‘llanilgan metodlar. Tadqiqot obyekti sifatida tabiiy tilni qayta ishlash tizimlari uchun mo‘ljallangan o‘zbek tilidagi raqamli matnlar korpusi hamda amaldagi o‘zbek yozuvi grafikasi (kirill va 1995-yilgi lotin alifbosi) tanlab olindi. Tadqiqotda tizimli tahlil, qiyosiy-grafemik tahlil, algoritmik modellashtirish va Unicode normalizatsiya (NFC) metodlaridan foydalanildi. Xususan, amaldagi lotin alifbosidagi apostrofli (*o‘, g‘*) va digrafli (*sh, ch, ng*) birliklarning tokenizatsiya hamda morfologik tahlil algoritmlariga ta’siri solishtirma tahlildan o‘tkazildi.

Olingan natijalar va ularning tahlili. Qiyosiy-grafemik va texnologik tahlillar natijasida keng ko‘lamli o‘zbek tili matnlar korpusini yaratishning to‘rt bosqichli funksional arxitekturasi ishlab chiqildi va amaldagi alifbo grafikasiga xos texnik muammolarning yechimlari tizimlashtirildi:

Korpus yaratish bosqichi	**Asosiy texnik va lingvistik jarayonlar**	**Amaldagi grafikaga xos murakkabliklar**	**Algoritmik yechim va qo‘llaniladigan standart**
1. Manbalarni yig‘ish (Data Collection)	Ochiq elektron kutubxonalar, OAV arxivlari, ilmiy va rasmiy hujjatlar bazasidan matnlarni yig‘ish	Kirill va lotin grafikasidagi matnlarning turli manbalarda aralash holda uchrashi	Ikki qatlamli saqlash: asl matnni arxivlash (Raw Layer) va yagona grafikaga o‘tkazish
2. Orfografik va Unicode normalizatsiya	HTML-teglar, takroriy matnlarni o‘chirish va harf-belgilarni yagona kodga keltirish	o‘, g‘ va tutuq belgisi (ъ) uchun 6 dan ortiq har xil Unicode apostrof belgilarining qo‘llanilishi	o‘, g‘ uchun U+02BB (modifier letter turned comma), tutuq belgisi uchun U+02BC standartini joriy etish
3. Kontekstual transliteratsiya	Kirill grafikadagi tarixiy va zamonaviy matnlarni avtomatik ravishda lotin alifbosiga o‘tkazish	*Ц*, *Щ*, *Ё*, *Ю*, *Я* harflari hamda *Е* harfining so‘z boshi va o‘rtasidagi turlicha o‘qilishi, *s+h* va *sh* omonimiyasi	Istisnolar lug‘ati va kontekstual qoidalarga asoslangan (rule-based) ikki yo‘nalishli konvertor
4. Tokenizatsiya va morfologik	Matnni tokenlarga ajratish, lemmatizatsiya	Agglyutinativlik (ko‘p affikslilik) hamda apostrof	Apostrofni so‘z ichki qismi sifatida tanuvchi

teglash	va so‘z turkumlarini (POS-tagging) aniqlash	belgisining so‘z o‘zagini ikkiga bo‘lib yuborishi	maxsus tokenizator va morfologik analizatorlar

Tahlillar shuni ko‘rsatadiki, amaldagi 1995-yilgi lotin alifbosiga to‘liq o‘tish va undan raqamli muhitda samarali foydalanish sharoitida korpus sifatini belgilovchi eng muhim omil — bu **Unicode belgilari unifikatsiyasi**dir. Amaliyotda foydalanuvchilar tomonidan klaviaturadagi oddiy tirnoq ('), teskari apostrof (`) yoki boshqa belgilardan foydalanilishi natijasida bitta so‘z (masalan, *O‘zbekiston*) kompyuter xotirasida bir necha xil kodlar ketma-ketligi sifatida saqlanadi. Bu esa tokenizatsiya jarayonida lug‘at hajmining sun’iy ravishda kengayib ketishiga va qidiruv tizimlarining matnni topa olmasligiga olib keladi. Shu bois, korpus konveyerining dastlabki bosqichidayoq barcha o‘ va g‘ harflari tarkibidagi belgilarni yagona U+02BB kodiga, tutuq belgisini esa U+02BC kodiga avtomatik konvertatsiya qiluvchi normalizatsiya filtrini o‘rnatish shart.

Muhokama. Olingan natijalar A. Po‘latov va N. Abdurahmonovalarning o‘zbek kompyuter lingvistikasida orfografik normalizatsiya va morfologik analizatorlar aniqligini oshirish zarurati haqidagi ilmiy xulosalarini to‘liq tasdiqlaydi [5, 121-b.; 6, 87-b.]. B. Mengliyev va Sh. Hamroyevalar ta’kidlaganidek, o‘zbek tili milliy korpusini yaratishda faqatgina hajm (tokenlar soni) emas, balki barcha funksional uslublar (rasmiy, ilmiy, badiiy, publitsistik va so‘zlashuv) muvozanatini ta’minlash asosiy mezon bo‘lishi kerak [7, 61-b.; 8, 90-b.]. D. Jurafskiy va J. Martinlar qayd etganidek, agglyutinativ va morfologik boy tillar uchun korpus tuzishda faqat statistik usullar yetarli bo‘lmay, tilning grammatik va morfofonologik qoidalarini o‘zida aks ettirgan gibrid (qoidaga asoslangan va neyron tarmoqli) algoritmlardan foydalanish yuqori samara beradi [2, 51-b.]. M. Sharipov va E. Kuriyozovlar ko‘rsatib o‘tganidek, ochiq kodli (open-source) korpuslar va standartlashtirilgan datasetlarning yaratilishi mahalliy tadqiqotchilar uchun milliy til modellarini ishlab chiqish imkoniyatini kengaytiradi [9, 138-b.].

Xulosa. O‘zbekiston Respublikasi Prezidentining PF–6084-son Farmoni, PQ–358-son va PQ–308-son qarorlari ijrosi doirasida tabiiy tilni qayta ishlash uchun keng ko‘lamli o‘zbek tilidagi matnlar korpusini yaratish milliy sun’iy intellekt texnologiyalarini rivojlantirishning fundamental asosi hisoblanadi. Tadqiqot doirasida ishlab chiqilgan to‘rt bosqichli korpus arxitekturasi kirill va lotin grafikalaridagi matnlarni yagona standart asosida birlashtirish imkonini beradi. O‘tkazilgan tahlillar asosida quyidagi ilmiy-amaliy tavsiyalar ilgari suriladi:

1. O‘zbek tilidagi barcha milliy va ixtisoslashgan NLP korpuslarida ma’lumotlarni saqlashning yagona kanonik standarti sifatida amaldagi lotin alifbosi asosidagi NFC (Unicode) formatini hamda o‘, g‘ (U+02BB) va tutuq belgisi (U+02BC) uchun qat’iy normalizatsiya protokollarini joriy etish.

2. Kirill grafikasida yaratilgan tarixiy, ilmiy va badiiy matnlar fondini yo‘qotmaslik maqsadida korpus konveyeriga kontekstual istisnolarni (*s+h* va *sh* birikmalarini, o‘zlashma so‘zlardagi tutuq belgisini) farqlay oladigan ikki yo‘nalishli transliteratsiya modulini integratsiya qilish.

3. Korpusning reprezentativligini ta’minlash uchun ma’lumotlar bazasini faqat OAV matnlari bilan cheklamay, ilmiy-texnik nashrlar, badiiy adabiyotlar, qonunchilik hujjatlari va og‘zaki nutq transkripsiyalari hisobidan mutanosib boyitib borish hamda ochiq kodli morfologik analizatorlarni rivojlantirish.

ADABIYOTLAR RO‘YXATI (REFERENCES):

1. Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991. – 179 p.
2. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. – 3rd ed. – Stanford: Stanford University, 2023. – 642 p.
3. McEnery T., Wilson A. Corpus Linguistics: An Introduction. – 2nd ed. – Edinburgh: Edinburgh University Press, 2001. – 235 p.
4. Mahmudov N., Madvaliyev A., Mahkamov N. O‘zbek tili imlosi va alifboni takomillashtirish masalalari // O‘zbek tili va adabiyoti. – Toshkent, 2021. – № 2. – B. 12-19.
5. Po‘latov A. K. Kompyuter lingvistikasi. – Toshkent: Akademnashr, 2011. – 520 b.
6. Abdurahmonova N. Z. Kompyuter lingvistikasi asoslari: O‘quv qo‘llanma. – Toshkent: Nodirabegim, 2021. – 216 b.
7. Hamroyeva Sh. M. O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari. – Buxoro: Durdona, 2018. – 148 b.
8. Mengliyev B. R., Karimov R. A. O‘zbek tili milliy korpusi: nazariya va amaliyot // O‘zbek tili va adabiyoti. – Toshkent, 2020. – № 3. – B. 84-90.
9. Sharipov M., Kuriyozov E. Uzbek Corpus: A Large-Scale Dataset for Natural Language Processing // Proceedings of the Language Resources and Evaluation Conference (LREC). – Marseille, 2022. – Pp. 132-139.
10. Elov B. B., Hamroyeva Sh. M. O‘zbek tili milliy korpusini yaratishda matnlarni avtomatik qayta ishlash va transliteratsiya masalalari // O‘zbekistonda xorijiy tillar. – Toshkent, 2022. – № 4 (45). – B. 38-46.